Project Proposal: Credit Card Default Prediction and Analysis

In the dynamic landscape of the financial sector, credit card companies face an ongoing challenge: accurately predicting and managing the risk of customer default. This project aims to address this challenge by developing a robust predictive model that assesses the likelihood of default payment for credit card clients in Taiwan, utilizing a dataset spanning from April 2005 to September 2005. The multifaceted nature of this problem is underscored by the complexities of demographic factors, credit information, and payment history.

Recent advancements in the financial industry have increasingly relied on predictive modeling as a cornerstone for evaluating credit risk. The pioneering work of Altman, as detailed in "Financial Ratios, Discriminant Analysis and the Prediction of Corporate Bankruptcy," highlights the significance of predictive analytics in identifying potential financial distress. Additionally, the exploration of ensemble learning techniques, illustrated by Breiman's "random forests", demonstrates their effectiveness in improving prediction accuracy, which is essential in the context of credit risk assessment. Understanding and navigating these complexities is pivotal for developing effective risk management strategies and informed decision-making by credit card companies.

The success of this project rests on two pivotal criteria: the accuracy and reliability of the predictive model and the depth of insights gained from exploratory data analysis. A successful outcome will empower credit card companies to make informed, data-driven decisions, reducing the incidence of defaults, and enhancing overall financial stability. The solution space is expansive, encompassing a multifaceted exploration of the dataset. This includes demographic analysis, feature engineering, and the development of a predictive model. The temporal aspect of the dataset, spanning a six-month period, provides a nuanced understanding of credit behavior, allowing for a comprehensive view of potential risk factors.

The dataset, sourced from the UCI Machine Learning Repository, comprises 25 variables, offering a rich tapestry of information. Articles such as "Credit Scoring and Its Applications" by Thomas and Edelman contribute valuable insights into the selection and utilization of credit-related features, guiding the incorporation of relevant variables into the analysis.

The project will commence with an in-depth exploration of the dataset, encompassing data cleaning, feature engineering, and demographic analysis. Leveraging methodologies from Hastie, Tibshirani, and Friedman's "The Elements of Statistical Learning," the predictive model will be developed and evaluated for accuracy and performance. The project's deliverables will include a meticulously documented and commented codebase, facilitating transparency and reproducibility. A comprehensive report will outline the methodology, findings, and recommendations, while a visually compelling slide deck will distill key insights for effective communication.

Future work will explore avenues for model optimization, consider additional data sources to enrich the analysis, and contemplate the development of an interactive dashboard. These considerations align with the evolving landscape of machine learning and data visualization, ensuring the project remains adaptable and relevant.